Nazir, S., Yousaf, M.H. y Velastín, S. A. (2018). Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Computers & Electrical Engineering*, 72, pp. 660-669.

DOI: https://doi.org/10.1016/j.compeleceng.2018.01.037

Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition

Saima Nazir [a], Muhammad Haroon Yousaf [a,*], Sergio A. Velastin [b,c]

[a] University of Engineering and Technology Taxila, Pakistan
[b] Universidad Carlos III de Madrid, Spain
[c] Queen Mary University of London, UK

A B S T R A C T

The detection of the spatial-temporal interest points has a key role in human action recognition algorithms. This research work aims to exploit the existing strength of bag-of-visual features and presents a method for automatic action recognition in realistic and complex scenarios. This paper provides a better feature representation by combining the benefit of both a well-known feature detector and descriptor i.e. the 3D Harris space-time interest point detector and the 3D Scale-Invariant Feature Transform descriptor. Finally, action videos are represented using a histogram of visual features by following the traditional bag-of-visual feature approach. Apart from video representation, a support vector machine (SVM) classifier is used for training and testing. A large number of experiments show the effectiveness of our method on existing benchmark datasets and shows state-of-the-art performance. This article reports 68.1% mean Average Precision (mAP), 94% and 91.8% average accuracy for Hollywood-2, UCF Sports and KTH datasets respectively.

Keywords:
Human action recognition
Local spatio-temporal features
Bag-of-visual features
Hollywood-2 dataset

1. Introduction

The impact of computer vision in areas such as video surveillance, human behavior analysis and human body motion has increased the interest in this field. This is mainly because of the large number of video data and the potentially enormous applications based on automatic video analysis. Under these circumstances, recognizing human actions in videos has enticed more and more interest of researchers and has been an active research area over the past three decades. The aim of human action recognition (HAR) is to recognize actions based on observation and environmental settings.

In recent years, a variety of work has been done to analyze and understand human actions, but due to the sheer complexity in identifying human actions, very few approaches have shown encouraging results in realistic environments. According to Niebles et al. [1], the most promising methods are based on advanced dimensionality reduction, bag of words, and random forest. Moreover, in recent years, methods such as dense trajectories [2] and hierarchical mined association rules [3] have shown promising results on realistic action datasets such as Hollywood-2 [4]. Amongst these techniques, bag of words (BoW) has been employed by many authors [1,4,5] for solving action recognition problems. The extensive use of BoW
is due to its capacity to address problems such as view independence, occlusions and scale invariance. Hence, instead of addressing these problems explicitly, improving recognition accuracy should be the focal point of research. This is particularly

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. G. Botella.
* Corresponding author.
E-mail address: haroon.yousaf@uettaxila.edu.pk (M.H. Yousaf).

important when dealing with videos where the camera is not static. Although satisfactory performance has been obtained in simple scenarios, many challenges remain when dealing with real 'life' videos which may include cluttered backgrounds, camera motions, occlusions, changes in viewpoints and temporal variations. Furthermore, the requirements of these techniques to deal with view independence, scale invariance, localization of actions and robustness of performance has escalated the problem.

Today, the recognition of human actions in simple scenarios with single view, no occlusion and unchanged background (e.g. KTH and Weizmann datasets) has been solved with high accuracy [6]. However, performance reduces significantly in realistic and complex scenarios e.g. the Hollywood-2 dataset. This paper contributes to the recognition of human actions in such scenarios. Our approach intends to utilize spatio-temporal domain information to provide a compact video representation for action recognition. We propose a new feature representation approach by using a popular spatio-temporal feature extractor and descriptor for human action recognition. To the best of our knowledge, this contribution is the first to use the proposed feature representation approach on human action recognition datasets. Representation of actions in the spatio-temporal domain has the advantage of being robust to camera motion, occlusion, background clutter, and scale and temporal variations.

The rest of the article is organized as follows. We first present a summary of relevant existing work of human action recognition approaches in Section 2, followed by a more detailed description of the proposed methodology and new feature representation approach in Section 3. Section 4 provides the experimental evaluation using three datasets, Hollywood-2, UCF Sports and KTH datasets for HAR in complex and realistic scenarios. Finally, Section 5 states the conclusions and future work.

2. Related work

Bag-of-visual features (BoVF) has been used by many authors as a mean for object and action recognition. This method, based on bag of words (BoW), was originally used in natural language processing for textual information retrieval. Many authors have used the BoW technique in computer vision to produce state-of-art results. BoW implementation requires selecting specific parameters for a sampling strategy (i.e. extracting localized features from a video sample), size of the code book, quantization, distance function used for in nearest-neighborhood assignment and choice of classifier.

As defined by Zelnik-Manor and Irani [7], actions can be considered as temporal objects which are usually spread over tens and hundreds of frames. A full body movement is considered as an action. A given action will have a number of variations. For example, the action of walking will differ according to the individual performing the action. The motive of human action recognition is to generalize over these variation [8]. A large number of research papers has been published to deal with HAR, some of which can be found in the surveys by Moeslund et al. in [9] and Poppe in [8].

In this section, we will provide a brief insight on strengths and weaknesses of promising papers that have employed BoW techniques or its variants for solving action recognition problem. Niebles et al. [1] focused on an unsupervised learning technique for human action categorization. They employed BoW representation for unsupervised learning of human actions by effective representation of sparse space time interest points [8] and probabilistic models which enables them to handle noise arising from camera motion and dynamic background. Their method localized and categorized multiple actions in a single video for a novel video sequence. In their paper they extracted spatial and spatial-temporal feature to categories human actions in a frame-by-frame basis. They had promising results, but their ideas were limited to few standard datasets such as KTH, Weizmann, figure skating actions and a complex video sequence [1] captured by a hand-held camera.

Further advancement was achieved by Liu and Shah [10] to discover the best number of clusters by using MMI (maximization of mutual information) in BoW pipeline, instead of the k-means clustering algorithm. They claim that, by exploiting the correlation of video-word clusters and spatial temporal pyramid matching, they were able to achieve rotation and translation invariance. After the introduction of more realistic and complex datasets such as Hollywood and Hollywood-2, the computer vision community needed to improve their techniques to overcome these new challenges. Papers by Marszalek et al. [4] and Laptev et al. [5] employed a standard BoW technique with combination of various feature descriptors to compare their results for each action. Their method did not take account of temporal information while computing SIFT, instead they exploited the contextual relationship between scene and action in the dataset to boost their results.

To further explore the challenges in real and complex environments, i.e. with cluttered backgrounds, changing viewpoints and occlusions, Wang et al. [11] compared and evaluated space-time features, using 4 different feature detectors and 6 different feature descriptors for the Hollywood-2 dataset. They applied a standard BoW technique to evaluate their results. They concluded that dense sampling is more difficult to handle than a relatively sparse number of interest points; however, it consistently outperforms all interest point detectors.

Gilbert et al. [3] were able to produce better results than previous authors in complex datasets such as Hollywood-2. They used an over complete feature set of dense 2D Harris corners which were hierarchically grouped together to form a complex structure. They utilized the Apriori algorithm of a data mining technique to handle the dense feature and learn from distinctive and descriptive association rules. Their idea was based on a proposed method by Quack et al. [12] for object detection using a frequent item set mining algorithm to generate discriminative association rules. These data mining techniques are fast and accurate in providing real time performance. Gilbert et al. [3] had to explicitly address the problem of scale invariance in their paper and their idea can localize the action in complex dataset such as Multi-KTH and Hollywood-2.

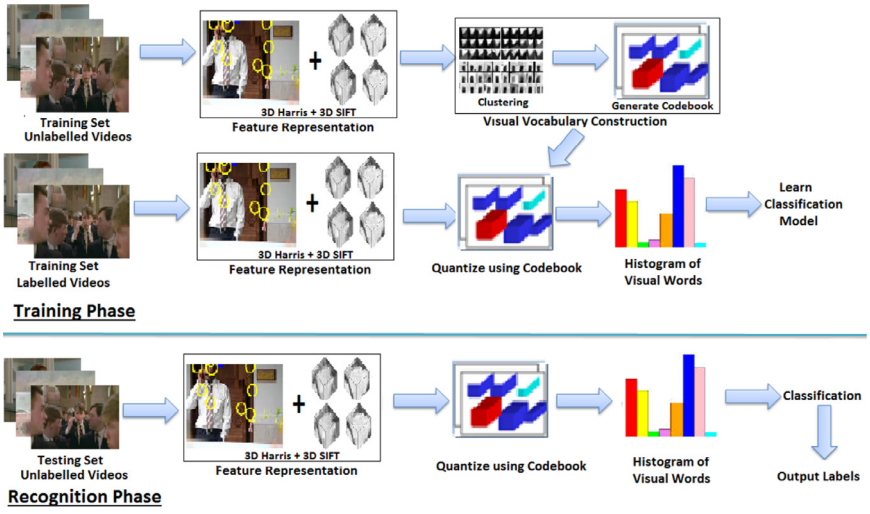

Fig. 1. Proposed methodology: bag of visual features model for human action recognition.

According to the authors, performance can be further improved with the help of more complex classifier architectures and exploitation of contextual relationship between action and scene provided in the Hollywood-2 dataset.

Another enhancement on the BoW technique was proposed by Ullah et al. [13]. To improve action recognition, they integrated additional non-local cues with BOW. They decompose each video into different region classes and augment the local features with the respective region-class label. They claim that segmentation method trained on additional training data will result in increasing the discrimination power of BOW approach by providing the additional supervision. They were able to improve average precision accuracy in complex dataset such as Hollywood-2.

Being influenced by the accomplishment of dense sampling in image classification and action recognition Wang et al. [2] used dense trajectories. They extracted sample dense points and obtained dense trajectories by tracking these points using optical flow field. They tested their technique on four standard benchmark datasets: YouTube, KTH, UCF Sports and Hollywood-2 and were able to produce better result than existing state of-art technique on most of these datasets. They claim that their method captures motion information efficiently and it is an adequate solution for the removal of camera motion by computing dense trajectories along with the motion boundaries descriptor.

3. Proposed methodology

In our work, incoming videos are assumed to be captured in realistic and uncontrolled environment and the input of the system and features were extracted on a frame by frame basis for further analysis. Our approach aims to address the existing problems in realistic and natural environments for which a currently available complex and realistic dataset (Hollywood-2) was used to train and test the proposed approach.

An unlabeled set of videos $VT = \{V_1, V_2, \ldots, V_n\}$ from the training video set was used to train the proposed system on a set of different action classes $L_1, L_2, \ldots, L_m$, where 'm' is the total number of a dataset's action classes. We obtained feature representation for a video V, by directly extracting space time interest points using 3DHarris and describing the extracted STIPs using 3D SIFT. Identifying the presence of multiple actions in a video sequence needs to be solved beforehand to deal with the frame-shot change problem present in the Hollywood-2 dataset. Manual annotation was performed to identify the presence of multiple actions. Each action was represented separately using start and end frame number in all these videos. The methodology adopted is a bag-of-visual features as shown in Fig. 1. The following sub-sections will discuss our proposed scheme in detail.

3.1. Local feature representation

Our proposed approach intends to highlight the importance and uniqueness of short range spatio-temporal visual pattern by employing local features. A sparse representation for each video sequence is generated by detecting the space time interest points. In contrast to Niebles et al. [1], where they have used Harris 3D corners (Spatio-Temporal Interest Points or STIPs) and a histogram of optical flow for feature representation, we used a novel feature representation technique using 3D Harris and 3D SIFT. Local interest points are extracted using the STIP detector 3D Harris proposed by Laptev [14] in the first step as shown in Fig. 2. It captures highly invariant information in space-time domain that leads to detection of interesting events even in complex and realistic scenarios.

3D Harris works on the concept of Harris and Forstner interest point operator. It detects local image regions that have a significant variation in spatio-temporal dimension. These interest points correspond to non-constant motion in spatio-

Fig. 2. Spatio-temporal interest points on Hollywood-2 dataset video samples.

temporal neighborhood. After this, small video patches were extracted using the detected STIPs and are described using 3D SIFT (3-Dimensional Scale-Invariant Feature Transform) descriptor proposed by Scovanner et al. in [15]. 3D SIFT, an extension
of SIFT in the 3D (spatio-temporal) domain, encodes local spatio-temporal information about the detected space time interest points. It allows robustness to noise and orientations. The length of a 3D SIFT descriptor is dependent on the number of sub-histogram as well as the number of bins used. In our approach a descriptor of 2560 elements was used to extract the spatio-temporal features [15].

As a result, each video V is represented by a set of spatial and temporal interest points as $V = (x_1, y_1, t_1, \alpha_1), (x_2, y_2, t_2, \alpha_2), \ldots, (x_n, y_n, t_n, \alpha_n)$ where n denotes the total number of detected STIPs, (x_i, y_i, t_i) denotes the space time position vector and α_i is the 3D SIFT feature representation of the i th detected space time interest point.

3.2. Visual vocabulary construction

In this step, a visual word codebook is formed by quantizing the obtained features using an unsupervised clustering algorithm, k-means clustering. For a set of local features $F = f_1, f_2, f_3, \ldots, f_m$ for all training videos, k-means clustering intends to partition the F set of feature points into k clusters $C = c_1, c_2, c_3, \ldots, c_k$. The center of each cluster, defines a spatio-temporal visual word and is a prototype associated with the kth cluster. Hence, each descriptor in a video will be associated to a nearest cluster in our visual word codebook by computing Euclidean distance, as it shows better result w.r.t other distance measures. The main aim of this algorithm is to minimize an objective function J as given in Eq. (1).

$$min(J) = \sum_{i=1}^{m} \sum_{j=1}^{k} ||F_i - C_j||^2 \tag{1}$$

where $||F_i - C_j||^2$ is the Euclidean distance measured between a feature vector F_i and a cluster center C_j. As a result, each video is represented as a set of spatio-temporal visual words from a visual words codebook. K-Means clustering is an efficient method but its final results are sensitive to initial values for the number of clusters. The impact of visual code book size is evaluated in our experiments and results are presented in the experimentation section.

3.3. Learning classification model

The occurrence of every word is counted to form a histogram of visual words. Every single feature vector f_i is mapped on a nearest center (visual word) C_x using Euclidean distance. Then any given video can be represented through a histogram (of visual words), $HW_j = \{w_1, w_2, w_3, \ldots, w_k\}$ where w_i is the occurrence frequency of ith visual word in video j and k is the total number of visual words. Action labels along with this histogram of visual words are then passed to a classifier for training.

For classification, supervised learning algorithms were implemented i.e. a Support Vector Machines (SVM) and a Naive Bayes Classifier. SVMs are the state-of-the-art large margin classifiers, which have gained popularity for human action recognition. Our classification model uses a multi-class learning model based on a binary support vector machine as mentioned in Eq. (2), where m is the number of unique action class labels.

$$m = \frac{m(m-1)}{2} \tag{2}$$

This multi-class learning model uses a one-versus-one coding design scheme. The SVM classifier uses John Platt's sequential minimal optimization algorithm for learning. It utilizes a polynomial kernel to derive a non-linear model to predict action labels from feature values.

During the testing phase, feature vectors are obtained by detecting and describing local interest point using 3D Harris and 3D SIFT for unlabeled videos, which are then quantized by a visual codebook created earlier. Histograms of visual words are created for the occurrence of every word which are passed on to a trained classifier to generate a targeted action class label.

We preferred the bag-of-visual features method due to its simplicity of representation and lack of pre-processing on input video to localize salient point. BOVF does not require segmentation or any other image processing task other than feature detection.

Fig. 3. Sample actions from Hollywood-2, KTH and UCF sports datasets.

4. Experimentation results and discussion

In this section the performance of our proposed methodology for both simple and realistic datasets, i.e. KTH, UCF Sports and Hollywood-2, was evaluated and compared with existing approaches for representation and classification of human actions. The impact of different parameters on the performance of our proposed algorithm was evaluated with respect to computational time and accuracy. Some sample video actions are shown in Fig. 3 for all Hollywood-2, KTH and UCF Sports datasets.

The Hollywood-2 (an extension of the Hollywood dataset) dataset is a large collection of natural dynamic human actions in diverse and realistic video settings. This dataset has been set as a benchmark for evaluating proposed methods for action recognition. As suggested by Marszalek et al. [4], performance is evaluated by calculating mean Average precision (mAP).

The UCF Sports dataset comprises 150 broadcast sports action video sequences captured in unconstrained environments. It is a challenging dataset captured in realistic environments with cluttered background, different viewpoints and occlusion. It contains 10 action classes i.e. Diving, Kicking, Walking, Lifting, SkateBoarding, RidingHorse,SwingBench, SideBench, GolfSwing and Running. We used average accuracy performance measure for action recognition as used by reported state-of-the-art methods.

KTH is a standard dataset for recognizing action in simple scenarios. It is captured in a controlled environment with simple background and has camera motion and zooming in few videos only. It has 6 action classes i.e. HandWaving, Walking, Boxing, Jogging, HandClapping and Running. Each action is performed by twenty five subjects in four different environments.

For Hollywood-2 we used the standard evaluation measure mean Average precision (mAP) as reported by other methods in literature. We used standard training and testing split, containing 823 videos and 884 videos respectively, for performance testing. For UCF Sports Leave-one-out Cross Validation method and average accuracy measure was used for comparison with state-of-the-art results. For the KTH Dataset, we have used the experimental setup defined in a state-of-the-art method in [4]. 16 subjects video sequences were used for training purposes while the rest of 9 subject's video sequences are used for testing.

Mean average precision is calculated using average precision (AP) for each action class. AP indicates the quality of ranked test videos according to the classification probability score. Average Precision is defined as:

$$AveragePrecision = \frac{1}{RV}\sum_{m}\frac{RV_m}{m}I_m \tag{3}$$

Where RV denotes the total number of true relevant videos and RV_m denotes the number of true relevant videos in the top m list.$I_m = 1$ if m^{th} video in the target dataset is relevant, otherwise 0. *Average Precision* $= 1$ when all RV are ranked on top of the irrelevant videos.

For feature vector description STIPs were described using 3D SIFT algorithm with 2560 dimension, followed by the generation of visual codebook. The setting of codebook size k is an important parameter for clustering. Fig. 4(a) and (b) show the histogram of visual word occurrence in training videos and testing videos respectively.

Table 1 shows the computation time for Hollywood-2 dataset by varying visual codebook size. As expected, computational time increases linearly w.r.t. increases in the visual codebook size. There was no significant change observed in recognition performance by varying codebook size.

In the recognition step, different classifiers i.e. SVM, Naive Bayes and KNN were trained for classification purpose. SVM and Naive Bayes show relatively better performance than KNN in terms of mAP for Hollywood2 Dataset, because they perform better in the presence of high dimensional data and outliers as compared to different lazy classifiers such as KNN, IBK
etc (Table 2).

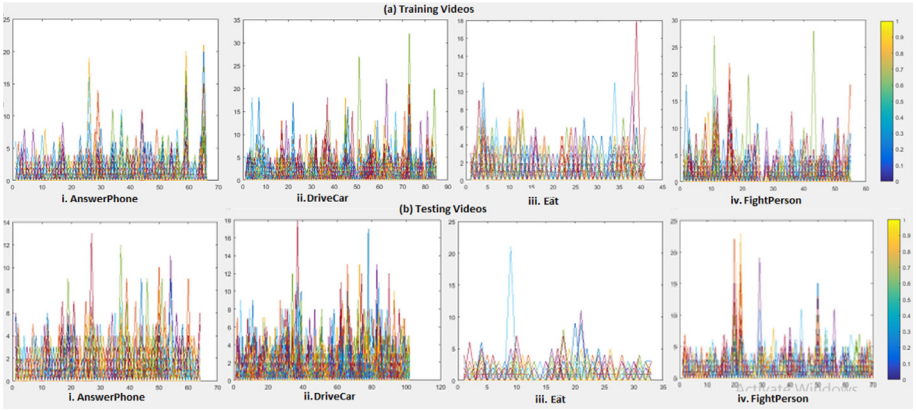

Fig. 4. Histograms of visual features for Hollywood 2 dataset (a) training videos (b) testing videos.

Table 1
Computational time and recognition performance for different visual code book size.

Visual code book size (k)	Computation time (mins)	Mean average precision (mAP)
K = 1000	20	66.7%
K = 2000	40	66.9%
K = 3000	60	66.0%
K = 4000	80	67.0%

Table 2
Performance of different classifiers for Hollywood-2 dataset.

Classifier	Performance (mAP)
SVM (with Normalization)	67.0%
SVM (without Normalization)	68.1%
Naive Bayes	63.0%

Table 3
Comparison with state-of-the-art methods for the Hollywood-2 dataset using average precision (AP) and mean average precision (mAP).

Action	Marszlek et al. [4]	Han et al. [16]	Gilbert et al. [3]	Ullah et al. [13]	Chakraborty et al. [17]	Wang et al. [2]	Ours
AnswerPhone	13.1	15.6	40.2	26.3	41.6	32.6	**61.5**
DriveCar	81.0	87.1	75.0	86.5	**88.5**	88	76.9
Eat	30.6	50.9	51.5	59.2	56.5	65.2	**69**
FightPerson	62.5	73.1	77.1	76.2	78.0	**81.4**	78.1
GetOutCar	8.6	27.2	45.6	45.7	47.7	52.7	**68.0**
HandShake	19.1	17.2	28.9	49.7	52.5	29.6	**63.1**
HugPerson	17.0	27.2	49.4	45.4	50.3	54.2	**66.4**
Kiss	57.6	42.9	56.5	59.0	57.4	65.8	**67.9**
Run	55.5	66.9	47.5	72.0	76.7	**82.1**	70.4
SitDown	30.0	41.6	62.0	62.0	62.5	62.5	**68.0**
SitUp	17.8	7.2	26.8	27.0	30.0	20.0	**56.7**
StandUp	33.5	48.6	50.7	58.8	60.0	65.2	**71**
mAP results	35.5	42.1	50.9	55.6	58.5	58.3	**68.1**

For the Hollywood-2 dataset, performance was computed using average precision for every action class and mAP to compare it with state-of-the-art reported results. Table 3 shows comparison with other methods. In our experiments, we have used the clean Hollywood-2 training dataset as defined by Laptev et al. [5] who constructed such a dataset using automatic training with action labels manually verified. Training and testing video sequences splits are as those proposed by other authors to allow direct comparison to their results. We argue that our approach performs better with respect to other approaches due to the capability of the selected feature representation techniques, as STIP performs better in environments with cluttered background, illumination changes and scale and view invariance. In 9 action class categories our approach reports best average class accuracy, thus results in better overall performance (mAP) so far reported in the literature.

The FightPerson class achieves the highest average class accuracy i.e. 78.1% followed by DriveCar, Standup and Run action class as 76.9%, 71.0% and 70.4% respectively. As STIPs also detect unwanted interest points in realistic environment that explains the effect on overall performance rate.

Table 4
Performance comparison for UCF sports dataset based on average accuracy per class.

Class	Yuan et al. [18]	Yao et al. [19]	Qiu et al. [20]	Zhu et al. [21]	Ours
Diving	100	100	86	100	96
Golf Swing	89	100	94	76	84
Kicking	100	73	75	80	87
Lifting	95	77	100	100	99
Riding Horse	58	43	92	75	95
Running	69	95	46	55	93
Skate Boarding	83	46	83	83	96
Swing Bench	100	91	100	90	97
Swing Side	77	100	100	95	95
Walking	91	92	59	91	93
Avg. Accuracy	86	82	84	85	94

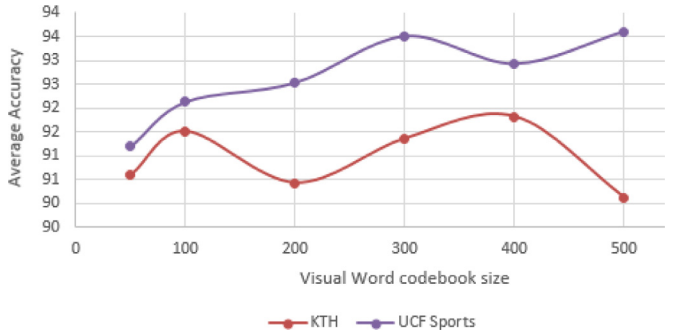

Fig. 5. Curve showing the influence of different vocabulary size on KTH and UCFSports datasets.

Table 5
Comparison with state-of-the-art work based on average accuracy per class for KTH dataset.

Class	Laptev et al. [5]	Liu et al. [10]	Neibles et al. [1]	Gilbert et al. [3]	Ours
Boxing	99	98	82	100	94.9
Handclapping	89	94	88	94	91.2
Handwaving	80	96	53	99	94.0
Jogging	97	89	93	91	91.2
Running	91	87	86	89	84.3
Walking	95	100	98	94	95.4

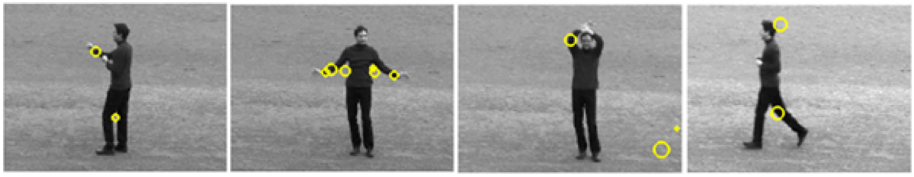

Fig. 6. Detected interest points for Boxing, HandClapping, HandWaving and Jogging actions for KTH dataset.

Table 4 reports results for the UCF Sports dataset, by setting visual codebook size to 500 visual words and using a Leave One out Cross Validation (LOOCV) method as used in the literature. Low accuracy is observed for Golf Swing and Kicking action classes, as similarities between two different actions classes can lead to confusion. Actions containing well-defined gestures can have large variations when performed in realistic scenarios. Therefore, dealing with the large variations of an action still remains a major challenge.

For the KTH dataset, we followed the standard approach for performance evaluation. 16 person video sequences were used to train a supervised classifier, a Support Vector Machine, and 9 person video sequences were used for testing. Fig. 5 shows average accuracy achieved by varying visual codebook size for KTH and UCF Sports dataset, with a visual codebook size of 50 and incrementing it to 500.

Table 5 shows each action class accuracy performance for KTH dataset with a visual vocabulary size of 400. As the KTH dataset is captured in a semi-controlled environment with homogeneous background, the number of detected interest points that have significant variation in local spatial-temporal neighborhood was quite low. In addition, when these interest points were further described by 3D SIFT they were thrown out due to their poor descriptive ability. As it can been seen in Fig. 6,

Table 6
Performance comparisons on Hollywood-2, UCF Sports and KTH. (mAP is used for Hollywood-2, and average accuracy for UCF Sports and KTH datasets).

Hollywood2		UCF sports		KTH	
Ullah et al. [13]	55.70%	Wang et al. [2]	88.20%	Tsai et al. [6]	100%
Wang et al. [2]	58.30%	Yuan et al. [18]	87.30%	Gilbert et al. [3]	94.50%
Jain et al. [22]	66.40%	Zhu et al. [21]	84.30%	Wang et al. [2]	94.20%
Sun et al. [23]	48.10%	Sun et al. [23]	86.60%	Sun et al. [23]	93.10%
Ours	68.10%	Ours	94.00%	Ours	91.82%

detected STIPs are well localized in both space and time domain but, in some cases, they are insufficient to differentiate events from each other and noise.

Since better performance is always obtained by encoding geometric information, this paper used widely used Space Time Interest Point detector 3D Harris and 3D SIFT for incorporating the information from both spatial and temporal domain. We believe that the main strength of our proposed approach that leads to improved results, is the use of 3D Harris. This results in well localized STIPs in the space and time domains and corresponds to meaningful events in a video. On the other hand, 3D SIFT is capable of handling challenges present in realistic scenarios such as occlusion, noise and dynamic background, as 3D SIFT encodes information in both spatial and temporal domains hence providing robustness to orientation and noise.

In Table 6 we have summarized state-of-the-art performance on the three different datasets. And get the highest mean average precision (mAP) for Hollywood2 and average accuracy for UCF Sports.

Our proposed method performs better than other methods by achieving 68.1% mAP for the Hollywood-2 dataset and 94% average accuracy for the UCF Sports dataset because of its robustness to existing challenges in realistic and complex scenarios. It also reports good results for a simple dataset such as KTH which is comparable to the state-of-the-art results, but it also highlights a limitation of interest points in some situations.

5. Conclusion

Better performance of the bag-of-visual feature approach is achieved for realistic and complex human action recognition datasets. It is shown that performance can be significantly improved in complex and realistic scenarios by incorporating spatio-temporal domain information to represent an action in the form of visual features. We have used a state-of-the art space time interest point detector and descriptor to capture the maximum possible information to represent an action. It represents video by utilizing characteristic shape and motion, independent of space time shifts. No prior segmentation like individual segmentation is needed for this approach. 3D Harris and 3D SIFT feature representation are capable to handle challenges present in realistic scenarios. Such feature representation approach provides robustness to noise and orientation and detects meaningful events. Our approach is general and shows better results on different type of human action recognition datasets. We have also performed comparison analysis with the state-of-art result for three different human action recognition datasets.

For future work, we can perform temporal segmentation to handle multiple actions in video sequences and exploit the co-occurrence based relations for visual words to increase the performance for the Hollywood-2 dataset. Another future direction is to incorporate the spatio-temporal contextual information that is ignored by the bag-of-visual feature approach. Instead of using handcrafted feature representation, we can also evaluate the strength of deep learning approach for recognizing human actions in uncontrolled environment.

Acknowledgments

Sergio A Velastin has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministe- rio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

References

[1] Niebles JC, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 2008;79(3):299–318.
[2] Wang H, Kläser A, Schmid C, Liu C-L. Action recognition by dense trajectories. In: Computer vision and pattern recognition (CVPR), 2011 IEEE confer- ence on. IEEE; 2011. p. 3169–76.
[3] Gilbert A, Illingworth J, Bowden R. Action recognition using mined hierarchical compound features. IEEE Trans Pattern Anal Mach Intell 2011;33(5):883–97.
[4] Marszalek M, Laptev I, Schmid C. Actions in context. In: Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on. IEEE; 2009.
p. 2929–36.
[5] Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on. IEEE; 2008. p. 1–8.
[6] Tsai D-M, Chiu W-Y, Lee M-H. Optical flow-motion history image (OF-MHI) for action recognition. Signal Image Video Process 2015;9(8):1897–906.
[7] Zelnik-Manor L, Irani M. Statistical analysis of dynamic actions. IEEE Trans Pattern Anal Mach Intell 2006;28(9):1530–5.

[8] Poppe R. A survey on vision-based human action recognition. Image Vis Comput 2010;28(6):976–90.
[9] Moeslund TB, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 2006;104(2):90–126.
[10] Liu J, Shah M. Learning human actions via information maximization. In: Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on. IEEE; 2008. p. 1–8.
[11] Wang H, Ullah MM, Klaser A, Laptev I, Schmid C. Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British machine vision conference. BMVA Press; 2009. p. 124–31.
[12] Quack T, Ferrari V, Leibe B, Van Gool L. Efficient mining of frequent and distinctive feature configurations. In: Computer vision, 2007. ICCV 2007. IEEE 11th international conference on. IEEE; 2007. p. 1–8.
[13] Ullah MM, Parizi SN, Laptev I. Improving bag-of-features action recognition with non-local cues.. In: BMVC, 10; 2010. p. 95–101.
[14] Laptev I. On space-time interest points. Int J Comput Vis 2005;64(2–3):107–23.
[15] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on multimedia. ACM; 2007. p. 357–60.
[16] Han D, Bo L, Sminchisescu C. Selection and context for action recognition. In: Computer vision, 2009 IEEE 12th international conference on. IEEE; 2009. p. 1933–40.
[17] Chakraborty B, Holte MB, Moeslund TB, Gonzalez J, Roca FX. A selective spatio-temporal interest point detector for human action recognition in complex scenes. In: Computer vision (ICCV), 2011 IEEE international conference on. IEEE; 2011. p. 1776–83.
[18] Yuan C, Li X, Hu W, Ling H, Maybank S. 3d R transform on spatio-temporal interest points for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 724–30.
[19] Yao A, Gall J, Van Gool L. A hough transform-based voting framework for action recognition. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. IEEE; 2010. p. 2061–8.
[20] Qiu Q, Jiang Z, Chellappa R. Sparse dictionary-based representation and recognition of action attributes. In: Computer vision (ICCV), 2011 IEEE inter- national conference on. IEEE; 2011. p. 707–14.
[21] Zhu Y, Zhao X, Fu Y, Liu Y. Sparse coding on local spatial-temporal volumes for human action recognition. In: Asian conference on computer vision. Springer; 2010. p. 660–71.
[22] Jain M, van Gemert JC, Snoek CG. What do 15,000 object categories tell us about classifying and localizing actions?. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 46–55.
[23] Sun L, Jia K, Chan T-H, Fang Y, Wang G, Yan S. DL-SFA: deeply-learned slow feature analysis for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 2625–32.

Saima Nazir is a Ph.D. Scholar at University of Engineering and Technology Taxila, Pakistan. Her field of work is 'Human Action Recognition in Complex and Realistic Scenarios'. She received the MS Software Engineering degree from UET Taxila Pakistan in 2014. Her main research interest includes Computer Vision and Machine learning.

Muhammad H. Yousaf is working as an Associate Professor in Computer Engineering Department at University of Engineering and Technology Taxila, Pakistan. He completed his Ph.D. Computer Engineering in 2012. His research interests includes Image Processing and Computer Vision. He has authored a number of papers in international conferences and journals. He is the recipient of BEST University Teacher Award by HEC Pakistan.

Sergio A. Velastin is UC3M Conex-Marie Curie Fellow in the Applied Artificial Intelligence Research Group at the Universidad Carlos III in Madrid. He previously worked as research professor at the University of Santiago de Chile and Professor in Applied Computer Vision at Kingston University, UK. He is also a Fellow of the Institution of Engineering and Technology (IET) and Senior Member of the IEEE.